



ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
Σχολή Επιστημών Υγείας
Πανεπιστήμιο Θεσσαλίας

Πρόγραμμα Μεταπτυχιακών Σπουδών (ΠΜΣ)
«Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και
Κλινική Βιοπληροφορική»

Διπλωματική εργασία με θέμα

Σύγκριση των μεθόδων συσταδοποίησης
Compare the types of clustering

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

Επιβλέπων καθηγητής: Ιωάννης Στεφανίδης

Χρυσούλα Δοξάνη

Ηλίας Ζιντζαράς

Κατερίνα-Μαρία Κοντούλη

Αύγουστος 2017

ΠΕΡΙΛΗΨΗ

Η διαδικασία συσταδοποίησης αποτελεί ένα πολύ χρήσιμο εργαλείο για την ομαδοποίηση δεδομένων, με σκοπό να σχηματίζονται δείγματα αντιπροσωπευτικά για το αρχικό σύνολο το οποίο βρίσκεται υπό μελέτη. Στην παρούσα διπλωματική εργασία γίνεται παρουσίαση των βασικότερων μεθόδων συσταδοποίησης. Έπειτα, αναλύονται οι διαδικασίες που έχουν αναφερθεί και δίνεται ιδιαίτερη βάση στους αλγορίθμους που χρησιμοποιούνται περισσότερο και είναι εύρεος διαδεδομένοι. Σε αυτές τις διαδικασίες συσταδοποίησης δίνονται και από ένα παράδειγμα υλοποίησης και ανάλυσης τους στο πρόγραμμα SPSS version 24. Τέλος, στο τελευταίο κεφάλαιο γίνεται μία σύνοψη όλων των αναφερόμενων μεθόδων και διευκρινίζεται ανάλογα την περίπτωση ποια μέθοδος κρίνεται ως καταλληλότερη.

ABSTRACT

The clustering process is a very useful tool for grouping data in order to form representative samples of the original data set which is under study. In this master's thesis the most popular methods of clustering are presented. Then, the methods are analyzed and the emphasis is given to the algorithms which are mostly used. An example of each of the most used clustering processes implementation and an analysis at the SPSS program version 24 is also given. At the last chapter a summary is provided of all the methods of clustering and it specifies which method is the most appropriate according to the data set.

ΕΙΣΑΓΩΓΗ

Η πρόοδος στην απόκτηση ψηφιακών δεδομένων και στην τεχνολογία αποθήκευσης τους, είχε σαν αποτέλεσμα την αύξηση του μεγέθους των βάσεων δεδομένων. Αυτό συνέβη σε όλους τους τομείς της ανθρώπινης ενασχόλησης. Ξεκινώντας από την καθημερινή ζωή όπως δεδομένα συναλλαγών, δεδομένα τηλεφωνικών κλήσεων και στατιστικά δεδομένα και φτάνοντας μέχρι και σε πιο εξεζητημένους τομείς όπως μοριακές βάσεις δεδομένων και ιατρικά αρχεία. Δεν αποτελεί έκπληξη λοιπόν το ενδιαφέρον να απομονώσουμε τα επιθυμητά δεδομένα και να εξάγουμε από αυτά πληροφορίες που μπορεί να είναι χρήσιμες.

Η συσταδοποίηση (clustering) είναι δύσκολο να οριστεί με ακριβείς όρους και μέσα σε συγκεκριμένο πλαίσιο, παρόλα αυτά ανήκει στην κατηγορία της μη επιβλεπόμενης μάθησης (Unsupervised learning) όπου μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής και στόχος είναι η χρήση κάποιου αλγορίθμου, ώστε αυτόματα να ανακαλύψουμε κάποια ενδιαφέρουσα δομή των δεδομένων.

Συσταδοποίηση είναι η διαδικασία ομαδοποίησης αντικειμένων με όμοια χαρακτηριστικά και η κατάταξη σε κλάσεις ή συστάδες ή συμπλέγματα έτσι ώστε να γίνει ευκολότερη η μελέτη και η εξαγωγή συμπερασμάτων από τα δεδομένα (Pang-Ning Tan, Mickael Steinbach, Vipin Kumar, 2006). Στην συσταδοποίηση οι συστάδες δεν είναι προκαθορισμένες αλλά προσδιορίζονται από τα δεδομένα. Το γεγονός αυτό είναι που αποτελεί και το δύσκολο σημείο στην διαδικασία μιας συσταδοποίησης.

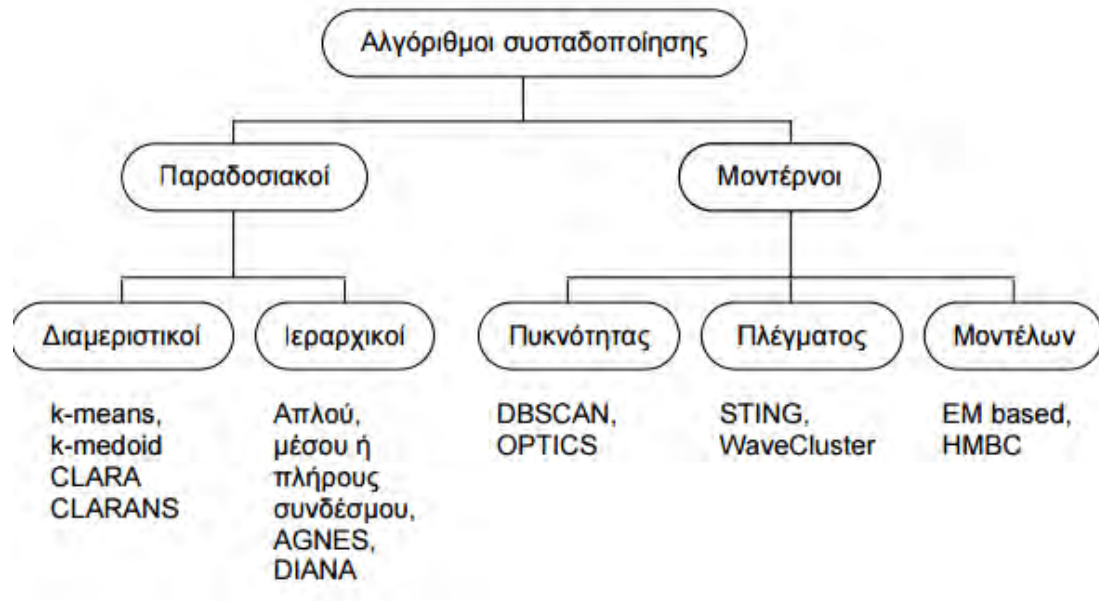
Πιο αναλυτικά, στο πρόβλημα της συσταδοποίησης μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις και χρειαζόμαστε κάποιον αλγόριθμο, ο οποίος θα ομαδοποιήσει αυτόματα τα δεδομένα σε συστάδες. Οι συστάδες που δημιουργούνται πρέπει να διαχωρίζουν ορθά τα δεδομένα. Αυτό πρακτικά σημαίνει ότι μια συστάδα θέλουμε να απαρτίζεται από αντικείμενα, όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της ίδιας συστάδας απ' ό,τι σε κάποιο άλλο αντικείμενο διαφορετικής συστάδας.

Η συσταδοποίηση έχει χρησιμοποιηθεί σε πολλά πεδία εφαρμογών, συμπεριλαμβανομένων της βιολογίας, ιατρικής, γενετικής, ανθρωπολογίας, μάρκετινγκ, οικονομίας, χρήσεων γης, ασφαλειών, ανάπτυξη πόλεων, αστρονομίας κλπ. Η συσταδοποίηση υλοποιείται με κάποιους αλγόριθμους, όπου ο καθένας δημιουργήθηκε για να εξυπηρετήσει διαφορετικές ανάγκες. Υπάρχει ένας μεγάλος αριθμός αλγορίθμων που ασχολούνται με την συσταδοποίηση, οι οποίοι έχουν μελετηθεί στην εξόρυξη γνώσης, την μηχανική μάθηση, την στατιστική και την αναγνώριση προτύπων.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η παρουσίαση των μεθόδων συσταδοποίησης, η ανάλυση και η κατανόηση του τρόπου υλοποίησης τους, η παρουσίαση των πλεονεκτημάτων και των μειονεκτημάτων τους και τέλος η πραγματοποίηση συγκρίσης μεταξύ των μεθόδων.

ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Όπως ήδη έχει αναφερθεί από την εισαγωγή με τον όρο συσταδοποίηση (clustering) εννοούμε την διαδικασία ομαδοποίησης αντικειμένων με όμοια χαρακτηριστικά και την κατάταξη σε κλάσεις ή συστάδες ή συμπλέγματα έτσι ώστε να γίνει ευκολότερη η μελέτη και η εξαγωγή συμπερασμάτων από τα δεδομένα (Pang-Ning Tan, Mickael Steinbach, Vipin Kumar, 2006). Σήμερα, έχει αναπτυχθεί ένας μεγάλος αριθμός τέτοιων μεθόδων για την ευκολότερη και καλύτερη συσταδοποίηση των δεδομένων. Η κάθε μέθοδος έχει τα δικά της πλεονεκτήματα και μειονεκτήματα και χρησιμοποιείται για δεδομένα διαφορετικής φύσης. Παρακάτω παρουσιάζονται οι κυριότερες κατηγορίες αλγορίθμων.



Όπως φαίνεται και από την εικόνα παραπάνω οι κύριες μέθοδοι συσταδοποίησης είναι 5, οι διαμεριστικοί, ιεραρχικοί, οι αλγόριθμοι πυκνότητας, οι πλέγματος και οι μοντέλων. Η κάθε κατηγορία από τις παραπάνω χωρίζεται σε υποκατηγορίες. Στην συνέχεια θα παρουσιαστούν τα κύρια στοιχεία της κάθε κατηγορίας.

- **Ιεραρχικοί αλγόριθμοι (Hierarchy algorithms):** δημιουργούν μία ιεραρχική δομή αποσύνθεσης του συνόλου δεδομένων σύμφωνα με κάποιο κριτήριο. Υπάρχει η διαιρετική και η συσσωρευτική προσέγγιση. Η ιεραρχική δομή υλοποιείται με την χρήση δέντροδιαγράμματος.
- **Διαμεριστικοί αλγόριθμοι (Partitioning algorithms):** κατασκευάζουν πολλές διαμερίσεις (partitions) των δεδομένων και τις αξιολογούν σύμφωνα με κάποιο κριτήριο απόστασης. Οι πιο γνωστοί είναι οι k-means και k-medoid.
- **Βασισμένοι σε πυκνότητα (Density-based):** βασίζονται στην συνεκτικότητα και συναρτήσεις πυκνότητας των σημείων των δεδομένων, έτσι οι συστάδες κατασκευάζονται σύμφωνα με κριτήρια πυκνότητας και συνεκτικότητας.
- **Βασισμένοι σε πλέγμα (Grid-based):** χρησιμοποιούν μία πολλαπλών επιπέδων υψηλής ανάλυσης κοκκοποιημένη δομή για να αναλύσουν τις συστάδες. Σε κάθε επίπεδο η ανάλυση μεγαλώνει. Παράδειγμα ο STING και ο Wave Cluster.

- **Βασισμένοι σε μοντέλα (Model-based):** υποθέτουν ένα μοντέλο για κάθε συστάδα και βρίσκουν το καλύτερο ταιρίασμα των δεδομένων σε αυτό διανέμοντας κάθε σημείο δεδομένου στη συστάδα στην οποία αναμένεται να έχει τη μεγαλύτερη πιθανότητα να ανήκει. Η εκτίμηση γίνεται μέσω της μέγιστης πιθανοφάνειας (maximum likelihood).

Στα επόμενα κεφάλαιο θα γίνει αναλυτική περιγραφή των μεθόδων που χρησιμοποιούνται συχνότερα, των ιεραρχικών μεθόδων, των διαμεριστικών αλγορίθμων και θα δοθεί έμφαση στην διαδικασία k-means και τέλος, θα αναλυθεί και η διαδικασία των αλγορίθμων συσταδοποίησης πυκνότητας και συγκεκριμένα την μεθόδου DBSCAN. Τέλος, θα γίνει αναφορά και στα υπόλοιπα μοντέλα συσταδοποίησης.

ΙΕΡΑΡΧΙΚΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗ (HIERARCHICAL CLUSTERING)

Οι ιεραρχικοί αλγόριθμοι συσταδοποίησης δημιουργούν μια ιεραρχία εμφωλιασμένων συσταδοποιήσεων. Δύο συστάδες λέγονται εμφωλιασμένες $E_p < E_q$ όταν ισχύει: $\forall x_i, x_j \in C_k$, κατά την συσταδοποίηση $E_p \rightarrow x_i, x_j \in C_k$ κατά την συσταδοποίηση E_q . Δηλαδή, συστάδες περιέχουν μεμονωμένα στοιχεία και άλλες συστάδες, οι οποίες με τη σειρά τους μπορεί να περιέχουν και αυτές άλλες, μικρότερες συστάδες, δημιουργώντας έτσι τα επίπεδα της ιεραρχίας (δεντροδιαγράμματα). Οι ιεραρχικοί αλγόριθμοι διακρίνονται σε δύο υποκατηγορίες: τους συσσωρευτικούς και τους διαιρετικούς. Οι αλγόριθμοι μπορούν να αναπαρασταθούν πλήρως με δενδροδιαγράμματα, τα οποία παρουσιάζουν τη διάταξη των συστάδων που δημιουργήθηκαν. Ουσιαστικά, κάθε επίπεδο ενός δενδροδιαγράμματος ορίζει ένα βήμα του αλγορίθμου. Το βασικό πλεονέκτημα των ιεραρχικών αλγορίθμων είναι ότι δεν χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό συστάδων, αφού οποιοσδήποτε αριθμός μπορεί να επιτευχθεί, απλά κόβοντας το δενδρόγραμμα στο κατάλληλο επίπεδο.

Στην διαδικασία υλοποίησης του ιεραρχικού αλγορίθμου σημαντικά σημεία που πρέπει να σημειωθούν είναι τα παρακάτω:

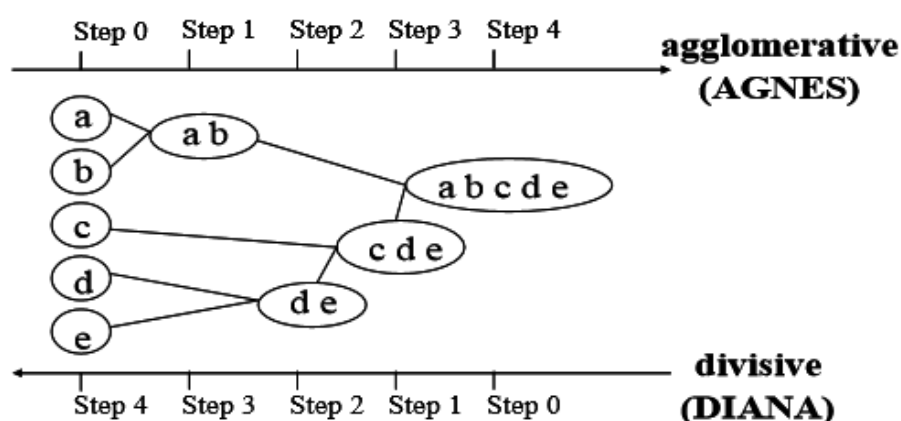
- Τα φύλλα του δέντρου αποτελούνται από ένα αντικείμενο.
- Κάθε αντικείμενο βρίσκεται σε μία συστάδα.
- Η ρίζα του δεντροδιαγράμματος περιέχει όλα τα αντικείμενα
- Οι εσωτερικοί κόμβοι παριστάνουν τις συστάδες που δημιουργούνται από την ένωση πολλών φύλλων ή και κόμβων-παιδιών.

Οι δύο κύριες προσεγγίσεις ιεραρχικών αλγορίθμων συσταδοποίησης είναι :

- **Συσσωρευτική μέθοδος (bottom up) :** στην αρχή κάθε αντικείμενο αποτελεί και μια συστάδα και μετά μικρές συστάδες ενώνονται σε μεγαλύτερες σε κάθε επίπεδο ιεραρχίας μέχρι το τελευταίο επίπεδο.
- **Διαιρετική μέθοδος (top down) :** είναι η αντίστροφη μέθοδος από την συσσωρευτική. Ξεκινά με ένα αντικείμενο το οποίο αποτελεί και μία συστάδα και έπειτα αυτή διασπάται σε όλο και μικρότερες συστάδες.

Οι περισσότεροι ιεραρχικοί αλγόριθμοι (π.χ. AGNES) ανήκουν στην συσσωρευτική κατηγορία. Διαφέρουν ως προς τον καθορισμό των μέτρων ομοιότητας μεταξύ των συστάδων.

Οι διαιρετικοί αλγόριθμοι (π.χ. DIANA) είναι σπανιότεροι και δεν απαιτούν αρχικό ορισμό των συστάδων αλλά χρειάζονται κριτήριο τερματισμού.



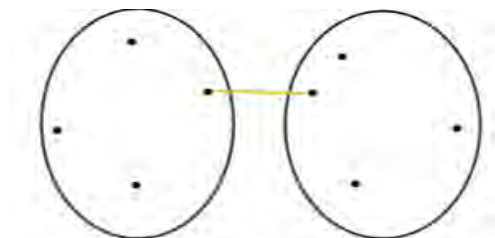
ΑΠΟΣΤΑΣΗ ΜΕΤΑΞΥ ΤΩΝ ΣΥΣΤΑΔΩΝ

Για την καλύτερη κατανόηση των συσσωρευτικών αλγορίθμων θα ήταν ορθό να γίνει μια ανάλυση ως προς τους τρόπους προσδιορισμού της απόστασης μεταξύ δύο συστάδων. Οι κυριότεροι τρόποι προσδιορισμού της απόστασης είναι:

- Ελάχιστης απόστασης ή απλού συνδέσμου (single link)
- Μέγιστης απόστασης ή πλήρους συνδέσμου (complete link)
- Μέσου όρου της συστάδας (group average)
- Απόσταση κεντρικών σημείων
- Μέθοδος του Ward

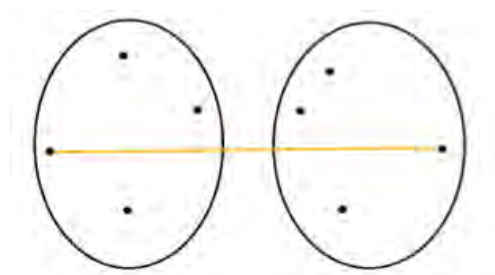
ΕΛΑΧΙΣΤΗΣ ΑΠΟΣΤΑΣΗΣ Ή ΑΠΛΟΥ ΣΥΝΔΕΣΜΟΥ

Ως απόσταση μεταξύ δύο συστάδων θεωρείται η μικρότερη απόσταση από κάθε μέλος της μίας συστάδας με κάθε μέλος της άλλης.



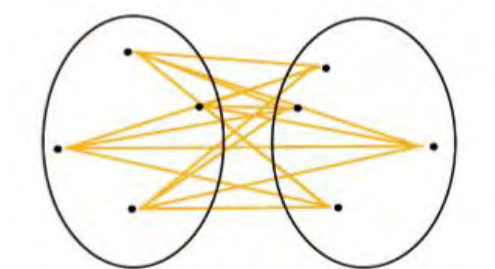
ΜΕΓΙΣΤΗΣ ΑΠΟΣΤΑΣΗΣ Ή ΠΛΗΡΟΥΣ ΣΥΝΔΕΣΜΟΥ

Ως απόσταση μεταξύ των συστάδων με την μέθοδο του πλήρους συνδέσμου θεωρείται η μεγαλύτερη απόσταση που δημιουργείται από τα μέλη των δύο συστάδων



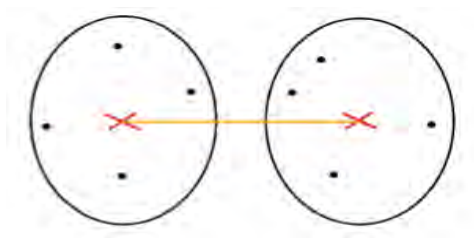
ΜΕΣΟΥ ΟΡΟΥ ΤΗΣ ΣΥΣΤΑΔΑΣ

Με την μέθοδο μέσου όρου ως απόσταση των συστάδων θεωρείται η μέση τιμή όλων των αποστάσεων που δημιουργούνται από κάθε πιθανό ζεύγος των δύο συστάδων.



ΑΠΟΣΤΑΣΗ ΚΕΝΤΡΙΚΩΝ ΣΗΜΕΙΩΝ

Ως απόσταση συστάδων με την μέθοδο των κεντρικών σημείων θεωρείται η απόσταση που σχηματίζεται από το κέντρο των συστάδων.



ΜΕΘΟΔΟΣ WARD

Τέλος, η βασική ιδέα πίσω από τη μέθοδο του Ward είναι ότι η απόσταση μεταξύ δύο συστάδων, C_i και C_j , είναι ίση με το πόσο θα αυξηθεί το άθροισμα των τετραγώνων της απόστασης των στοιχείων της κάθε συστάδας από το αντίστοιχο κεντροειδές (της κάθε συστάδας) μετά τη συγχώνευσή τους, C_{ij} , δηλαδή:

$$D_W(C_i, C_j) = \sum_{x \in C_i} (x - r_i)^2 + \sum_{x \in C_j} (x - r_j)^2 - \sum_{x \in C_{ij}} (x - r_{ij})^2$$

όπου r_i είναι το κεντροειδές της συστάδας C_i , r_j είναι το κεντροειδές της συστάδας C_j , και r_{ij} είναι το κεντροειδές της συστάδας C_{ij} , που προκύπτει από τη συγχώνευσή τους. Πρόκειται για το ιεραρχικό ανάλογο του k-means.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΩΝ ΙΕΡΑΡΧΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ

Γενικά οι αλγόριθμοι που ανήκουν σε αυτήν την κατηγορία χρησιμοποιούνται συνήθως όταν τα αντικείμενα που πρέπει να συσταδοποιηθούν απαιτούν ιεραρχία. Επίσης, μέσα από μελέτες έχει προκύψει ότι αυτού του είδους η συσταδοποίηση παράγει καλύτερης ποιότητας συστάδες. Επιπλέον, δεν απαιτεί αρχική δήλωση του αριθμού των συστάδων που θα δημιουργηθούν. Ωστόσο, οι συσσωρευτικοί ιεραρχικοί αλγόριθμοι είναι ακριβοί όσον αφορά την ένταση των υπολογισμών και των απαιτήσεων αποθήκευσης. Το γεγονός ότι όλες οι συγχωνεύσεις είναι τελικές μπορεί να προκαλέσει προβλήματα για θορυβώδη, μεγάλης κλίμακας δεδομένα, όπως τα δεδομένα εγγραφών. Παρόλα αυτά, τα δύο αυτά προβλήματα της αποθήκευσης και των θορυβωδών σημείων μπορούν να ξεπεραστούν χρησιμοποιώντας την μέθοδο αλγορίθμων k-means που ανήκει στους διαμεριστικούς αλγορίθμους.

ΑΛΛΕΣ ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ

Η μέθοδος **BIRCH** εισήγαγε την έννοια της συσταδοποίησης αντικειμένων αλλά και του CF-δέντρου. Αρχικά διαχωρίζει τα αντικείμενα ιεραρχικά χρησιμοποιώντας την δομή ενός CF-δέντρου. Σταδιακά προσαρμόζει την ποιότητα των συστάδων. Η μέθοδος BIRCH

είναι κατάλληλη για μεγάλες βάσεις δεδομένων, αλλά δυστυχώς δεν αποδίδει καλά εάν οι συστάδες δεν είναι σφαιρικού σχήματος ή εάν έχουν μεγάλες διαφοροποιήσεις ως προς το μέγεθος. (Tian Zhang, Raghu Ramakrishnan, Miron Livny, 1997)

Η μέθοδος **CURE** χρησιμοποιεί ένα σύνολο από σημεία, αντί να χρησιμοποιεί μόνο ένα σημείο για να εκπροσωπεί μία συστάδα σύμφωνα με τις μεθόδους που είναι βασισμένες σε κέντρα βάρους. Αυτός ο σταθερός αριθμός των αντιπροσωπευτικών σημείων της συστάδας επιλέγεται έτσι ώστε να είναι καλά διεσπαρμένα και μετά να μειώνονται ως προς το κέντρο βάρους της κάθε συστάδας σύμφωνα με τον παράγοντα μείωσης. Μετά οι συστάδες συγχωνεύονται επανειλημμένα βάση της ομοιότητάς τους. Η ομοιότητα μεταξύ δύο συστάδων μετριέται με την ομοιότητά του πιο κοντινού ζεύγους από τα αντιπροσωπευτικά σημεία που ανήκουν σε διαφορετικές συστάδες. Η μέθοδος CURE αγνοεί τις πληροφορίες αλληλοσυνεκτικότητας των αντικειμένων μέσα στην συστάδα.

Η μέθοδος **ROCK** ενεργεί πάνω σε έναν παραγόμενο γράφημα ομοιοτήτων. Αντί να χρησιμοποιεί την έννοια της απόστασης για να μετρήσει την ομοιότητα μεταξύ των σημείων, χρησιμοποιείται η έννοια των δεσμών η οποία περιέχει περισσότερες γενικές πληροφορίες του διαστήματος συστάδων έναντι του μέτρου ομοιότητας απόστασης που εξετάζει μόνο την τοπική απόσταση μεταξύ δύο σημείων. Το πρόβλημα του αλγορίθμου είναι ότι δεν είναι επιτυχής στο να κανονικοποιεί δεσμούς συστάδων. Το αποτέλεσμα του ROCK δεν είναι καλό για πολύπλοκες συστάδες με ποικίλες πυκνότητες δεδομένων. Ακόμα είναι πολύ ευπαθής στην επιλογή των παραμέτρων και ευαίσθητος στον θόρυβο.

K-MEAN ΜΕΘΟΔΟΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Η διαδικασία διαμεριστικών αλγορίθμων δημιουργεί μία κατανομή ενός επιπέδου από το σύνολο των αντικειμένων που διατίθενται για συσταδοποίηση. Υπάρχει ένας μεγάλος αριθμός τέτοιων μεθόδων για την υλοποίηση αυτής της τεχνικής, παρόλα αυτά οι δύο πιο διακεκριμένες είναι η K-means και η K-medoid.

Η μέθοδος k-means ορίζει μία πρωτότυπη συστάδα με βάση κάποιο στοιχείο, το κεντροειδές, το οποίο συνήθως αποτελεί το μέσο από μία ομάδα σημείων. Από την άλλη μεριά η μέθοδος k-medoid ορίζει μία συστάδα με βάση το στοιχείο το οποίο αποτελεί αντιπροσωπευτικό δείγμα για όλη την συστάδα. Οι διαφορές των δύο μεθόδων βασίζονται σε δύο βασικές λεπτομέρειες οι οποίες εξάγονται εξ' ορισμού.

- Αρχικά, το βασικό στοιχείο για την δημιουργία της συστάδας με την μέθοδο k-means δεν αποτελεί ένα πραγματικό σημείο, αλλά αυτό προκύπτει από τον υπολογισμό του

μέσου όλων των σημείων. Αντίθετα, με την k-medoid που το στοιχείο αυτό είναι πραγματικό.

- Επιπλέον, η μέθοδος k-means εφαρμόζεται σε n-διάστατους συνεχείς χώρους, ενώ η μέθοδος k-medoid εφαρμόζεται σε ένα ευρύ φάσμα δεδομένων, αφού απαιτεί μόνο ένα μέτρο για τον υπολογισμό της απόστασης δύο σημείων.

Στην συνέχεια θα γίνει ανάλυση για την μέθοδο k-means, διότι αποτελεί την πιο παλιά αλλά και την πιο διαδεδομένη μέθοδο συσταδοποίησης.

ΒΑΣΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΤΗΣ ΜΕΘΟΔΟΥ K-MEANS

Τα βασικά βήματα του αλγόριθμου είναι τα εξής:

- Επιλογή του αριθμού των ομάδων.
- Τυχαία δημιουργία k ομάδων και ορισμός των κεντροειδών των ομάδων.
- Μεταβίβαση του κάθε σημείου στο κεντροειδές της κοντινότερης ομάδας.
- Υπολογισμός των νέων κεντροειδών των ομάδων.
- Επανάληψη μέχρι να συγκλίνει ο αλγόριθμος σε κάποιο κριτήριο.

Ο αλγόριθμος k-means ξεκινάει με k σημεία τα οποία είτε προσδιορίζονται τυχαία, είτε χρησιμοποιώντας κάποια ερευνητικά δεδομένα, τα οποία ονομάζονται κεντροειδή της συστάδας και δηλώνουν το κέντρο βάρους της συστάδας ($\text{Centroid} = \sum_{i=1}^n p_i / N$). Το k υποδηλώνει πόσες συστάδες θέλουμε ο αλγόριθμος να δημιουργήσει. Ο αλγόριθμος εκτελεί επαναληπτικά δύο βήματα. Το πρώτο βήμα αφορά την ανάθεση σε κάποια συστάδα, ενώ το δεύτερο βήμα αφορά τον επαναπροσδιορισμό και τη μετατόπιση του κεντροειδούς κάθε συστάδας.

Πιο αναλυτικά, όσον αφορά στο πρώτο βήμα, δηλαδή την ανάθεση σε κάποια συστάδα, ο αλγόριθμος εξετάζει κάθε δείγμα σε σχέση με τα κεντροειδή των συστάδων. Με χρήση κάποιου μέτρου απόστασης, αναθέτει το εξεταζόμενο δείγμα στη συστάδα, της οποίας το κεντροειδές είναι το πλησιέστερο ως προς το συγκεκριμένο δείγμα. Στο δεύτερο βήμα, παίρνοντας τον μέσο όρο των δειγμάτων κάθε συστάδας, επανυπολογίζονται τα κεντροειδή της κάθε συστάδας, ώστε το κεντροειδές να είναι πιο αντιπροσωπευτικό στην πρόσφατα διαμορφωμένη συστάδα. Ο αλγόριθμος εκτελεί επαναληπτικά αυτά τα δύο βήματα, έως ότου τα κεντροειδή των συστάδων να μετατοπίζονται ελάχιστα και σε απόσταση μικρότερη από κάποια δοθείσα τιμή κατωφλίου. Ως εναλλακτικό κριτήριο τερματισμού του αλγορίθμου μπορεί να χρησιμοποιηθεί και ο αριθμός επαναλήψεων του αλγορίθμου.

Για να οριστεί ποιο κεντροειδές είναι το κοντινότερο σε κάθε σημείο θα πρέπει να οριστεί σε κάθε εφαρμογή της μεθόδου εξ'αρχής με ποια διαδικασία θα υπολογίζεται η

απόσταση των σημείων. Οι τρόποι που χρησιμοποιούνται συνήθως είναι η μέθοδος της ευκλείδειας απόστασης, η οποία χρησιμοποιείται κυρίως στον χώρο R^2 , και η μέθοδος Manhattan που χρησιμοποιείται κυρίως στον χώρο R .

Η μέθοδος της ευκλείδειας απόστασης χρησιμοποιεί για τον υπολογισμό της απόστασης δύο σημείων, $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_n)$ την 2-νορμική απόσταση.

τύπος της 2-νορμικής απόστασης : $d_2 = (\sum_{i=1}^n |x_i - y_i|)^{1/2}$

Η 1-νορμική απόσταση ονομάζεται και νορμική ταξί ή απόσταση Manhattan, επειδή είναι για παράδειγμα η απόσταση που διανύει ένα αυτοκίνητο σε μια πόλη που ορίζεται από οικοδομικά τετράγωνα (εάν δεν υπάρχουν μονόδρομοι).

Τυπος για την μέθοδο Manhattan (1-νορμική απόσταση) : $d_1 = \sum_{i=1}^n |x_i - y_i|$.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ K-MEANS

Γενικότερα, η μέθοδος k-means είναι απλή και μπορεί να χρησιμοποιηθεί σε ένα ευρύ φάσμα δεδομένων. Επίσης, είναι αρκετά αποδοτικοί παρόλο που ο αλγόριθμος μπορεί να εκτελείται αρκετές φορές. Παρόλα αυτά, η μέθοδος k-means δεν είναι κατάλληλη για μη σφαιρικά δεδομένα και δεδομένα διαφορετικών μεγεθών και διαφορετικών πυκνοτήτων. Επιπλέον, όταν τα δεδομένα περιέχουν απομακρυσμένα σημεία δημιουργείται πρόβλημα στην σωστή υλοποίηση του k-means αλγορίθμου. Η ανίχνευση των απομακρυσμένων σημείων και η απομάκρυνση τους βοηθά σημαντικά στην καλύτερη υλοποίηση του αλγορίθμου. Τέλος, η k-means μέθοδος περιορίζεται όταν υπάρχει από τα δεδομένα μία ένδειξη για το κέντρο.

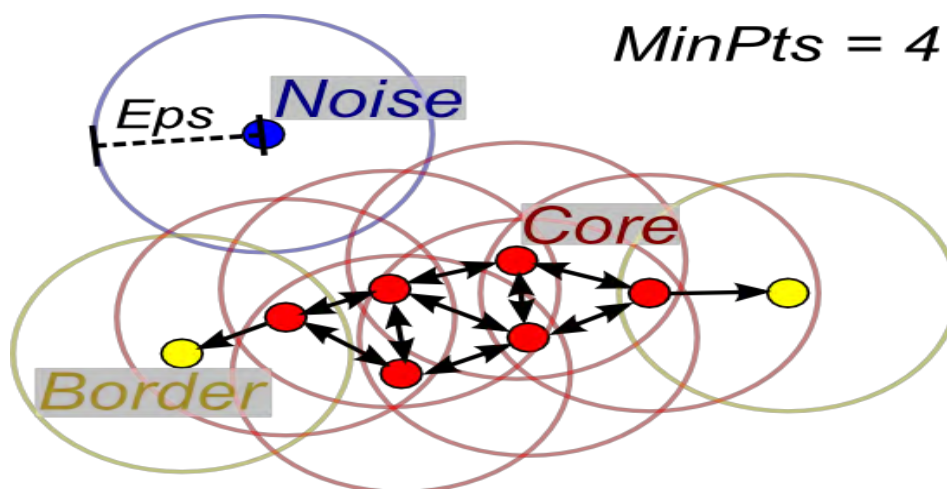
Αναλυτικότερα, ένα από τα μειονεκτήματα του αλγορίθμου k-means είναι το γεγονός ότι δεν υπάρχει κάποιος αυτοματοποιημένος τρόπος επιλογής του k, δηλαδή του αριθμού των συστάδων. Ο αριθμός των συστάδων δίνεται ως είσοδος από τον χρήστη και η επιλογή του σωστού αριθμού αφήνεται στη δική του γνώση και εμπειρία. Συνεπώς, η διαδικασία επιλογής του αριθμού συστάδων, ενδεχομένως, να απαιτήσει την εξερεύνηση και μελέτη των δεδομένων, για παράδειγμα, μέσα από οπτικοποιήσεις, προκειμένου να καταλήξουμε στον σωστό αριθμό συστάδων. Τέλος, πρέπει να σημειωθεί ότι είναι κατάλληλος ο αλγόριθμος k-means και γενικότερα όλοι οι αλγόριθμοι της κατηγορίας αυτής μόνο για σφαιρικές συστάδες. Αυτό είναι πραγματικά ένα γενικό πρόβλημα για όλες τις μεθόδους που χρησιμοποιούν διαμερισμό επειδή χρησιμοποιούν μόνο ένα κέντρο βάρους για να αντιπροσωπεύσουν μια συστάδα, και η συγκέντρωση όλων των άλλων σημείων αποφασίζεται από τη σχετική κοντινότητα τους στα κέντρα βάρους των συστάδων. Οι περισσότερες μέθοδοι που χρησιμοποιούν διαμερισμό έχουν παρόμοια αποτελέσματα.

ΜΕΘΟΔΟΙ ΒΑΣΙΣΜΕΝΕΣ ΣΤΗΝ ΠΥΚΝΟΤΗΤΑ

Η πιο σημαντική μέθοδος σε αυτήν την κατηγορία είναι η DBSCAN. Ο αλγόριθμος DBSCAN {Density-Based Spatial Clustering of Applications with Noise} δημιουργεί συστάδες με ένα ελάχιστο μέγεθος και πυκνότητα. Ως πυκνότητα ορίζεται το ελάχιστο πλήθος σημείων που απέχουν συγκεκριμένη απόσταση μεταξύ τους.

Έστω ένα σύνολο από σημεία στον χώρο, τα οποία θέλουμε να συσταδοποιήσουμε. Στο πλαίσιο της συσταδοποίησης με τον αλγόριθμο DBSCAN, τα σημεία αυτά κατηγοριοποιούνται σε κεντρικά (core points), προσεγγίσιμα ή πυκνά-προσεγγίσιμα (density-reachable points) ή απομακρυσμένα (outliers) με βάση τους παρακάτω κανόνες:

- Ένα σημείο p είναι κεντρικό σημείο, αν τουλάχιστον $MinPts$ σημεία βρίσκονται σε απόσταση ϵ από αυτό και αυτά τα σημεία είναι άμεσα προσεγγίσιμα από το p . Κανένα σημείο δεν είναι προσεγγίσιμο από μη κεντρικό σημείο.
- Ένα σημείο q είναι πυκνά-προσεγγίσιμο από ένα σημείο p , αν υπάρχει μονοπάτι p_1, \dots, p_n , με $p_1 = p$ και $p_n = q$, όπου κάθε p_{i+1} είναι άμεσα προσεγγίσιμο από το p_i , δηλαδή όλα τα σημεία του μονοπατιού πρέπει να είναι κεντρικά, με πιθανή εξαίρεση το q .
- Σημεία, τα οποία δεν είναι προσεγγίσιμα από κανένα άλλο σημείο, είναι ακραία.



Σε αυτό το διάγραμμα, έχει οριστεί $\text{MinPts}=4$. Όλα τα κόκκινα σημεία είναι κεντρικά σημεία, επειδή η περιοχή που περιβάλλει αυτά τα σημεία σε μία ακτίνα ϵ , περιέχει τουλάχιστον 4 σημεία. Επειδή είναι όλα προσεγγίσιμα το ένα από το άλλο, σχηματίζουν όλα μαζί μία συστάδα. Τα κίτρινα σημεία δεν είναι κεντρικά σημεία αλλά είναι προσβάσιμα από το σημείο A (μέσω άλλων σημείων), και έτσι ανήκουν στην συστάδα. Το μπλε σημείο N είναι ένα θορυβώδες σημείο το οποίο δεν είναι ούτε ένα κεντρικό σημείο ούτε ένα απευθείας προσβάσιμο σημείο.

Αν το σημείο p είναι κεντρικό σημείο, τότε σχηματίζει μια συστάδα μαζί με όλα τα σημεία (κεντρικά ή μη), τα οποία είναι προσεγγίσιμα από αυτό. Κάθε συστάδα περιέχει τουλάχιστον ένα κεντρικό σημείο. Η προσεγγισιμότητα δεν είναι συμμετρική σχέση, καθώς εξ ορισμού μπορεί κανένα σημείο να μην είναι προσεγγίσιμο από ένα μη κεντρικό σημείο, ανεξάρτητα από την απόσταση μεταξύ τους. Δηλαδή, ένα μη κεντρικό σημείο μπορεί να είναι προσεγγίσιμο, αλλά κανένα σημείο να μην μπορεί να προσεγγιστεί από αυτό. Συνεπώς, μια επιπλέον έννοια διασύνδεσης απαιτείται για τον ορισμό των συστάδων που δημιουργεί ο DBSCAN. Δυο σημεία p και q είναι πυκνά-συνδεδεμένα, αν υπάρχει σημείο s , τέτοιο ώστε τα p και q να είναι (πυκνά-)προσεγγίσιμα από το σημείο s . Η πυκνή-συνδεσιμότητα είναι συμμετρική σχέση. Έτσι, μια συστάδα εμφανίζει τις δυο ακόλουθες ιδιότητες:

- Όλα τα σημεία μιας συστάδας είναι αμοιβαία πυκνά-συνδεδεμένα μεταξύ τους.
- Αν ένα σημείο είναι πυκνά-προσεγγίσιμο από κάποιο σημείο της συστάδας, τότε είναι και αυτό, επίσης, σημείο της συστάδας.

Ο αλγόριθμος DBSCAN χρησιμοποιεί δυο παραμέτρους, το ϵ και το MinPts , δηλαδή τον ελάχιστο αριθμό σημείων που απαιτούνται για τη δημιουργία μιας πυκνής περιοχής. Ξεκινάει από ένα τυχαίο σημείο, το οποίο δεν έχει επισκεφθεί. Για το επιλεγμένο σημείο ανακτώνται τα σημεία στην ϵ -γειτονιά, δηλαδή σημεία που απέχουν το πολύ απόσταση ϵ από το επιλεγμένο σημείο. Αν υπάρχει επαρκής αριθμός σημείων, δηλαδή μεγαλύτερος του MinPts , ο αλγόριθμος δημιουργεί μια συστάδα. Σε αντίθετη περίπτωση, το σημείο μαρκάρεται προσωρινά ως θόρυβος. Μαρκάζονται προσωρινά, διότι σημεία που έχουν μαρκαιστεί ως θόρυβος, μπορεί στην πορεία να βρεθούν σε κάποια ϵ -γειτονιά άλλου επιλεγμένου σημείου και έτσι τελικά να γίνουν μέρος κάποιας συστάδας. Αν ένα σημείο αποτελεί πυκνό τμήμα μιας συστάδας, τότε σίγουρα η ϵ -γειτονιά του είναι υποσύνολο της συστάδας αυτής. Έτσι, όλα τα σημεία μέσα στην ϵ -γειτονιά προστίθενται στη συστάδα, καθώς επίσης και τα σημεία στην ϵ -γειτονιά καθενός από αυτά τα σημεία. Αυτή η διαδικασία συνεχίζεται μέχρι να ολοκληρωθεί η εύρεση της πυκνά-συνδεδεμένης συστάδας. Έπειτα ένα νέο σημείο επιλέγεται και ακολουθείται η παραπάνω διαδικασία, για την ανακάλυψη επιπλέον συστάδας ή θορύβου.

Σημεία, τα οποία μάρκαρε ο αλγόριθμος ως θόρυβο και που τελικά δεν έγιναν μέρος κάποιας συστάδας αποτελούν ακραία σημεία.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ DBSCAN

Ο αλγόριθμος DBSCAN, σε αντίθεση με άλλους αλγορίθμους όπως ο k-means που αναλύθηκε παραπάνω, δεν απαιτεί εκ των προτέρων προσδιορισμό των συστάδων. Το σχήμα των συστάδων που δημιουργούνται είναι αυθαίρετο και δεν έχει κάποιο συγκεκριμένο σχήμα όπως σφαιρικό. Για παράδειγμα, μπορεί να εντοπίσει μία συστάδα η οποία εκτείνεται γύρω από μία άλλη συστάδα. Το γεγονός αυτό συμβαίνει λόγω της παραμέτρου MinPts, η οποία ελαττώνει το φαινόμενο της αλυσίδας των συστάδων. Το φαινόμενο της αλυσίδας των συστάδων συμβαίνει όταν διαφορετικές συστάδες συνδέονται με μια λεπτή γραμμή σημείων. Επιπλέον, ένα ακόμα θετικό στοιχείο για τον αλγόριθμο DBSCAN είναι ότι χαρακτηρίζεται από καλή ευαισθησία στον θόρυβο και δεν επηρεάζεται από ακραίες τιμές. Για την υλοποίηση του αλγορίθμου χρειάζονται μόνο δύο παράμετροι, (MinPts, ϵ), όπου μετά από σωστή αξιολόγηση και παρατήρηση των δεδομένων ο προσδιορισμός τους κρίνεται εύκολος. Παρόλο που ο αλγόριθμος DBSCAN έχει αρκετά πλεονεκτήματα, διαθέτει και αυτός κάποια μειονεκτήματα. Αρχικά, δεν είναι απόλυτα ντετερμινιστικός, υπό την έννοια ότι τα περιθωριακά σημεία μιας συστάδας μπορεί είτε να ανήκουν σε αυτήν, είτε να ανήκουν σε κάποια γειτονικά, ανάλογα με την σειρά επεξεργασίας. Αυτή η περίπτωση όμως έχει μικρές πιθανότητες εμφάνισης και γενικότερα μικρό αντίκτυπο σαν αποτέλεσμα. Επιπλέον, η ποιότητα των αποτελεσμάτων εξαρτάται από τη μετρική απόστασης που θα χρησιμοποιηθεί. Η πιο κοινή μετρική απόστασης είναι η Ευκλείδεια απόσταση. Ειδικά όμως, για πολυδιάστατα δεδομένα, η συγκεκριμένη μετρική είναι σχεδόν ανούσια, κάνοντας έτσι δύσκολη την επιλογή της παραμέτρου ϵ . Ωστόσο, αυτό μπορεί να συμβεί σε οποιονδήποτε αλγόριθμο χρησιμοποιεί την Ευκλείδεια απόσταση. Ένα ακόμα σημαντικό μείνέκτημα στην χρήση του αλγορίθμου DBSCAN είναι ότι δεν μπορεί να συσταδοποιήσει καλά σύνολα από δεδομένα με μεγάλες διαφορές πυκνότητας, καθώς δεν μπορεί να εντοπιστεί κάποιος συνδυασμός MinPts- ϵ , που να είναι κατάλληλος για όλες τις συστάδες.

GRID-BASED ΑΛΓΟΡΙΘΜΟΙ

Οι μέθοδοι που είναι βασισμένες σε πλέγμα, αρχικά κβαντικοποιούν τον χώρο συσταδοποίησης σε έναν πεπερασμένο αριθμό κελιών και έπειτα εφαρμόζουν την μέθοδο της συσταδοποίησης στα κελιά που έχουν προηγουμένως δημιουργηθεί. Το κύριο πλεονέκτημα

των μεθόδων αυτών είναι πως η ταχύτητά τους εξαρτάται μόνο στην ανάλυση του πλέγματος και όχι στο μέγεθος της βάσης δεδομένων. Τέτοιες μέθοδοι είναι κατάλληλες περισσότερο για σύνολα δεδομένων υψηλής πυκνότητας με μεγάλο πλήθος αντικειμένων σε μικρό χώρο. (Guonjun Gan, Chaoqun Ma, Jianhong Wu, 2007)

MODEL-BASED ΑΛΓΟΡΙΘΜΟΙ

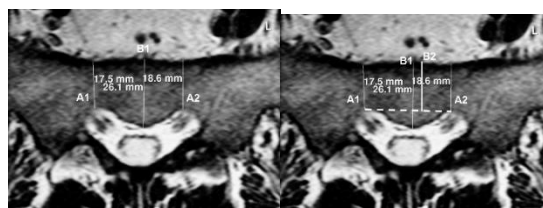
Προσπαθούν να βελτιστοποιήσουν το ταίριασμα μεταξύ των δεδομένων και ενός μαθηματικού μοντέλου που θα τα περιγράφει. Ο Αλγόριθμος EM (Expectation, Maximization) δημιουργεί συστάδες διανέμοντας με επαναληπτική επανατοποθέτηση, κάθε σημείο δεδομένου, στη συστάδα στην οποία αναμένεται να έχει τη μεγαλύτερη πιθανότητα να ανήκει. Ο EM βρίσκει την εκτίμηση μέγιστης πιθανοφάνειας (maximum likelihood) για μία παράμετρο 'μέσου'. Ο Hierarchical Model-based clustering (HMBC) συνενώνει ζεύγη συστάδων που αναλογούν στην ελάχιστη μείωση της μεταξύ τους πιθανοφάνειας.

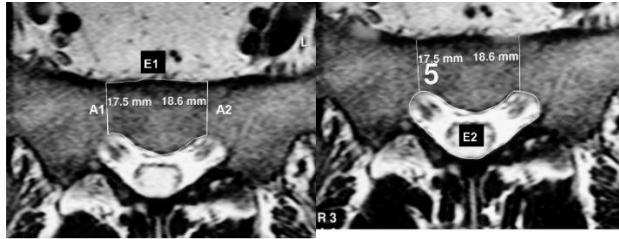
ΠΑΡΑΔΕΙΓΜΑΤΑ

Σε αυτό το σημείο θα αναλύσουμε με παραδείγματα το πως υλοποιούνται οι πιο συνήθεις τρόποι συσταδοποίησης οι οποίοι είναι η ιεραρχικής συσταδοποίησης και η μέθοδος k-means. Τα δύο παραδείγματα θα πραγματοποιηθούν στο spss πρόγραμμα version 24. Παρακάτω, παρουσιάζεται το αρχικό πρόβλημα και στην συνέχεια γίνεται η ανάλυση τους.

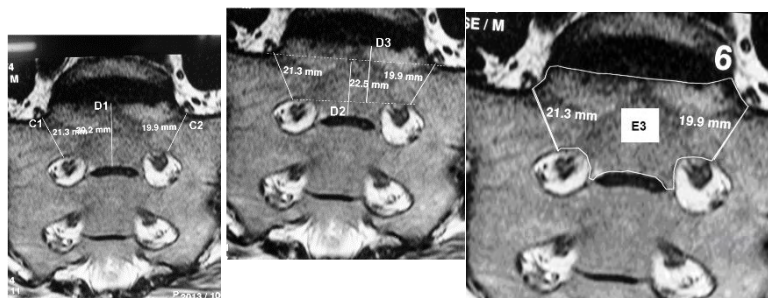
ΠΡΟΒΛΗΜΑ

12 morphometric measurements of sacrum were taken from 20 individuals. The measured variables are: A1, A2, B1, B2, E1, E2 (denoted: Level-1), G1, G2, D1, D2, D3, E3 (denoted Level-2) (see X-rays). Identify groups of individuals with similar morphological sacrum and interpret the results.





Transverse plane, level 1: A1 - anteroposterior (εμπροσπίσθια) diameter of right pedicle (εγκάρσια απόφυση), A2 - anteroposterior diameter of left pedicle, B1 - anteroposterior diameter of vertebra body, B2 - line between anterior aspect of vertebra body and the level of the lateral recess, E1 - area of vertebra body, E2 - area of spinal canal.



Coronal plane, level 2: C1 - superior-inferior diameter of right pedicle (κάθετη απόφυση), C2 - superior-inferior diameter of left pedicle, D1 - superior-inferior diameter of S1 vertebra body, D2 - line between the level of the ala and the inferior aspect of vertebra body, D3 - line between superior aspect of vertebra body and the level of roof of the foramen, and E3 - area of vertebra body (σώμα σπονδύλου).

A/A	A1	A2	B1	B2	E1	E2	G1	G2	D1	D2	D3	E3
1	20.2	21.2	37.2	29.2	1382.6	332.9	19.2	17.2	29.2	19.2	22.2	1522.3
2	21.3	21.9	26.0	15.1	772.9	310.8	18.8	18.2	26.9	18.1	26.4	1608.8
3	19.9	19.9	19.3	19.3	717.3	423.2	21.0	23.6	20.0	19.7	19.2	1172.9
4	15.2	16.6	27.2	20.0	753.5	461.2	21.5	22.4	27.6	23.4	24.3	1504.3
5	11.1	12.0	23.4	16.1	583.0	397.0	14.1	13.9	23.0	12.9	19.9	1295.8
6	8.2	11.2	21.3	13.9	666.0	473.8	14.8	18.4	22.3	10.7	17.3	1252.7
7	12.5	12.0	20.5	15.1	557.5	424.3	17.4	15.9	25.0	21.0	19.2	817.8
8	9.8	9.5	21.0	12.4	415.3	436.6	16.7	16.3	25.4	18.1	17.5	1049.8
9	13.7	13.5	21.4	16.2	575.2	331.1	17.4	18.9	25.2	19.5	22.8	1078.5
10	16.6	16.4	23.8	18.7	657.9	582.6	23.9	21.0	25.3	17.2	21.1	1114.1
11	17.2	15.3	23.0	17.6	700.5	436.2	17.6	15.0	28.4	18.9	25.8	1399.7
12	26.2	14.8	16.3	18.5	735.6	496.6	17.1	18.1	28.3	15.6	27.2	1687.5
13	19.3	21.0	18.8	16.4	782.6	439.4	13.4	15.2	20.6	9.9	18.7	1100.9
14	16.3	16.8	19.4	17.3	614.8	380.1	18.5	17.4	27.0	19.1	23.2	1395.5
15	15.2	16.5	22.1	16.7	547.2	303.5	18.0	16.5	25.7	18.1	21.7	1282.6
16	12.7	13.6	20.2	15.0	583.1	329.0	17.8	16.3	27.4	13.2	24.3	1394.6
17	14.2	17.1	24.8	18.4	733.5	352.0	16.5	17.4	28.4	16.8	227.1	1513.0
18	15.6	16.5	23.3	19.2	578.0	418.8	18.4	17.5	24.6	19.3	19.9	1313.6
19	23.0	18.4	28.7	20.5	1116.7	478.5	21.8	24.0	25.7	21.3	22.9	1679.1
20	16.2	17.2	21.9	17.6	607.5	476.1	18.1	18.5	27.6	16.6	24.1	1489.1

(Ζιντζαράς, 2017)

HIERARHICAL CLUSTERING

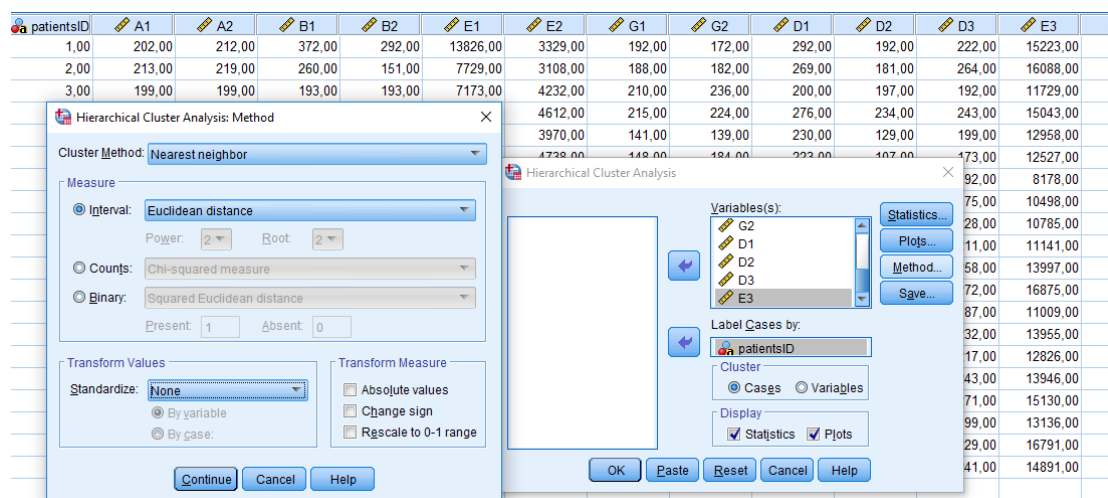
Μέσω του spss θα κάνουμε αρχικά συσταδοποίηση με την μέθοδο **hierarchical clustering**.

Έχουμε εισάγει τα δεδομένα στο spss

patientsID	A1	A2	B1	B2	E1	E2	G1	G2	D1	D2	D3	E3
1.00	202.00	212.00	372.00	292.00	13826.00	3329.00	192.00	172.00	292.00	192.00	222.00	15223.00
2.00	213.00	219.00	260.00	151.00	7729.00	3108.00	188.00	182.00	269.00	181.00	264.00	16088.00
3.00	199.00	199.00	193.00	193.00	7173.00	4232.00	210.00	236.00	200.00	197.00	192.00	11729.00
4.00	152.00	166.00	272.00	200.00	7535.00	4612.00	215.00	224.00	276.00	234.00	243.00	15043.00
5.00	111.00	120.00	234.00	161.00	5830.00	3970.00	141.00	139.00	230.00	129.00	199.00	12958.00
6.00	82.00	112.00	213.00	139.00	6660.00	4738.00	148.00	184.00	223.00	107.00	173.00	12527.00
7.00	125.00	120.00	205.00	151.00	5575.00	4243.00	174.00	159.00	250.00	210.00	192.00	8178.00
8.00	98.00	95.00	210.00	124.00	4153.00	4366.00	167.00	163.00	254.00	181.00	175.00	10498.00
9.00	137.00	135.00	214.00	162.00	5752.00	3311.00	174.00	189.00	252.00	195.00	228.00	10785.00
10.00	166.00	164.00	238.00	187.00	6579.00	5826.00	239.00	210.00	253.00	172.00	211.00	11141.00
1.00	172.00	153.00	230.00	176.00	7005.00	4362.00	176.00	150.00	284.00	189.00	258.00	13997.00
12.00	262.00	148.00	163.00	185.00	7356.00	4966.00	171.00	181.00	283.00	156.00	272.00	16875.00
13.00	193.00	210.00	188.00	164.00	7826.00	4394.00	134.00	152.00	206.00	99.00	187.00	11009.00
14.00	163.00	168.00	194.00	173.00	6148.00	3801.00	185.00	174.00	270.00	191.00	232.00	13955.00
15.00	152.00	165.00	221.00	167.00	5472.00	3035.00	180.00	165.00	257.00	181.00	217.00	12826.00
16.00	127.00	136.00	202.00	150.00	5831.00	3290.00	178.00	163.00	274.00	132.00	243.00	13946.00
17.00	142.00	171.00	248.00	184.00	7335.00	3520.00	165.00	174.00	284.00	168.00	2271.00	15130.00
18.00	156.00	165.00	233.00	192.00	5780.00	4188.00	184.00	175.00	246.00	193.00	199.00	13136.00
19.00	230.00	184.00	287.00	205.00	11167.00	4785.00	218.00	240.00	257.00	213.00	229.00	16791.00
20.00	162.00	172.00	219.00	176.00	6075.00	4761.00	181.00	185.00	276.00	166.00	241.00	14891.00

Και ακολουθώντας τα βήματα που φαίνονται παρακάτω θα κάνουμε την συσταδοποίηση.

Classify → hierarchical clustering



Η μέθοδος (Cluster Method) που έχει επιλεγεί είναι η nearest neighbours analysis (or single linkage analysis) η οποία επιτρέπει την ταυτοποίηση ομάδων(συστάδων) αντικειμένων, στην προκειμένη περίπτωση ατόμων με κοινά μορφολογικά χαρακτηριστικά στο ιερό οστό, με βάση τις μεταβλητές που εισάγονται.

Η μέθοδος ξεκινά μετρώντας τις αποστάσεις μεταξύ των αντικειμένων, στην προκειμένη περίπτωση ατόμων. Στην αρχή, το κάθε άτομο αποτελεί μια ομάδα από μόνο του, και σταδιακά με βάση τη μετρούμενη Ευκλείδεια απόσταση (όπως φαίνεται και η επιλογή της μεθόδου μέτρησης στην αμέσως προηγούμενη εικόνα) ξεκινάει η δημιουργία μεγαλύτερων ομάδων καθώς η μετρούμενη απόσταση μεγαλώνει. Στην αρχή δηλαδή προστίθενται στην ίδια ομάδα τα δύο άτομα που απέχουν λιγότερο. Στη συνέχεια τα δύο άτομα με τη δεύτερη μεγαλύτερη απόσταση σχηματίζουν ομάδα και η διαδικασία αυτή συνεχίζεται μέχρι τελικά να τοποθετηθούν και τα 20 άτομα που περιλαμβάνονται στο συγκεκριμένο παράδειγμα. Τελικά δηλαδή, όλα τα άτομα κατατάσσονται στην ίδια ομάδα. Αυτή η διαδικασία φαίνεται απεικονιστικά στο δένδrogramma.

Στον επόμενο πίνακα (proximity ή distance matrix) φαίνονται οι υπολογιζόμενες αποστάσεις μεταξύ των ατόμων που περιλαμβάνονται στη μελέτη. Τα άτομα υπάρχουν τόσο οριζόντια όσο και κάθετα και δίνεται η αντίστοιχη τιμή της απόστασης εκεί που διασταυρώνονται. Προφανώς για το ίδιο άτομο η απόσταση είναι 0.

Proximity Matrix

	Euclidean Distance																				
Case	1: 1,00	2: 2,00	3: 3,00	4: 4,00	5: 5,00	6: 6,00	7: 7,00	8: 8,00	9: 9,00	10: 10,00	11: 1,00	12: 12,00	13: 13,00	14: 14,00	15: 15,00	16: 16,00	17: 17,00	18: 18,00	19: 19,00	20: 20,00	
1: 1,00		,000	6164,868	7572,429	6425,258	8339,191	7790,687	10890,917	10819,028	9216,286	8686,485	7009,730	6880,189	7413,729	7799,502	8698,624	8100,456	6812,511	8358,704	3416,463	7891,799
2: 2,00	6164,868		,000	4538,044	1845,683	3765,585	4065,565	8277,906	6757,160	5664,684	5761,630	2545,197	2056,658	5242,682	2746,022	3968,771	2871,353	2297,903	3700,384	3890,784	2628,481
3: 3,00	7572,429	4538,044		,000	3358,205	1849,099	1093,673	3897,030	3269,621	1943,057	1802,527	2283,304	5203,903	997,283	2491,009	2354,679	2762,816	4055,014	1983,234	6472,774	3391,066
4: 4,00	6425,258	1845,683	3358,205		,000	2775,059	2675,201	7150,208	5673,357	4797,236	4197,741	1204,050	1883,844	4056,518	1943,599	3416,172	2424,692	2315,857	2627,914	4035,440	1479,098
5: 5,00	8339,191	3765,585	1849,099	2775,059		,000	1212,270	4795,580	3004,762	2274,394	2707,406	1621,790	4324,848	2825,098	1067,773	1014,731	1202,509	3389,516	306,948	6624,764	2106,399
6: 6,00	7790,687	4065,565	1093,673	2675,201	1212,270		,000	4511,227	3247,814	2431,409	1771,868	1565,998	4415,445	1951,195	1790,175	2102,252	2193,129	3624,158	1212,127	6209,464	2440,217
7: 7,00	10890,917	8277,906	3897,030	7150,208	4795,580	4511,227		,000	2724,506	2774,779	3508,264	5994,198	8908,752	3624,014	5822,756	4804,022	5852,618	7502,128	4963,162	10285,071	6752,283
8: 8,00	10819,028	6757,160	3269,621	5673,357	3004,762	3247,814	2724,506		,000	1939,203	2907,479	4516,300	7164,648	3713,159	4033,158	2990,424	3984,090	6058,336	3106,846	9435,249	4813,223
9: 9,00	9216,286	5664,684	1943,057	4797,236	2274,394	2431,409	2774,779	1939,203		,000	2672,874	3605,141	6513,321	2355,836	3232,461	2079,041	3162,967	5060,382	2510,142	8221,317	4366,913
10: 10,00	8686,485	5761,630	1802,527	4197,741	2707,406	1771,868	3508,264	2907,479	2672,874		,000	3239,271	5852,139	1910,873	3494,630	3443,963	3856,526	5104,251	2703,055	7352,993	3931,508
11: 1,00	7009,730	2545,197	2283,304	1204,050	1621,790	1565,998	5994,198	4516,300	3605,141	3239,271		,000	2964,096	3103,192	1026,683	2342,153	1592,979	2481,224	1509,476	5032,696	1351,315
12: 12,00	6880,189	2056,658	5203,903	1883,844	4324,848	4415,445	8908,752	7164,648	6513,321	5852,139	2964,096		,000	5914,924	3370,079	4867,582	3706,296	3025,685	4134,480	3820,091	2373,660
13: 13,00	7413,729	5242,682	997,283	4056,518	2825,098	1951,195	3624,014	3713,159	2355,836	1910,873	3103,192	5914,924		,000	3444,808	3272,155	3721,013	4727,691	2961,839	6692,711	4276,649
14: 14,00	7799,502	2746,022	2491,009	1943,599	1067,773	1790,175	5822,756	4033,158	3232,461	3494,630	1026,683	3370,079	3444,808		,000	1523,111	606,953	2651,544	979,574	5849,961	1343,337
15: 15,00	8698,624	3968,771	2354,679	3416,172	1014,731	2102,252	4804,022	2990,424	2079,041	3443,963	2342,153	4867,582	3272,155	1523,111		,000	1205,735	3638,088	1233,638	7158,019	2758,388
16: 16,00	8100,456	2871,353	2762,816	2424,692	1202,509	2193,129	5852,618	3984,090	3162,967	3856,526	1592,979	3706,296	3721,013	606,953	1205,735		,000	2799,279	1214,962	6232,149	1766,789
17: 17,00	6812,511	2297,903	4055,014	2315,857	3389,516	3624,158	7502,128	6058,336	5060,382	5104,251	2481,224	3025,685	4727,691	2651,544	3638,088	2799,279		,000	3337,128	4820,055	2703,230
18: 18,00	8358,704	3700,384	1983,234	2627,914	306,948	1212,127	4963,162	3106,846	2510,142	2703,055	1509,476	4134,480	2961,839	979,574	1233,638	1214,962	3337,128		,000	6538,419	1870,673
19: 19,00	3416,463	3890,784	6472,774	4035,440	6624,764	6209,464	10285,071	9435,249	8221,317	7352,993	5032,696	3820,091	6692,711	5849,961	7158,019	6232,149	4820,055	6538,419		,000	5436,578
20: 20,00	7891,799	2628,481	3391,066	1479,098	2106,399	2440,217	6752,283	4813,223	4366,913	3931,508	1351,315	2373,660	4276,649	1343,337	2758,388	1766,789	2703,230	1870,673	5436,578		,000

This is a dissimilarity matrix

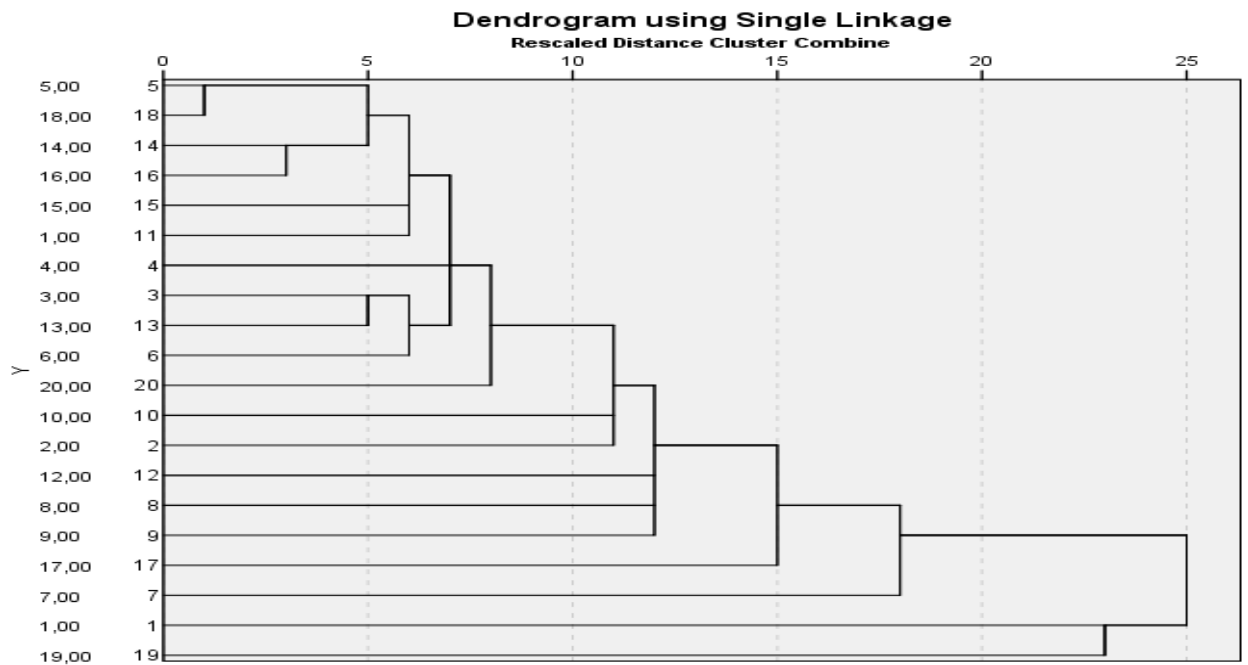
Στον πίνακα που ακολουθεί, φαίνονται πιο απλοποιημένα τα στοιχεία του προηγούμενου. Δηλαδή έχει γίνει κατάταξη των ατόμων με βάση την απόσταση, ξεκινώντας από την μικρότερη. Η ελάχιστη απόσταση είναι μεταξύ των ατόμων 5 και 18. Οπότε σχηματίζεται μια ομάδα. Η αμέσως επόμενη είναι μεταξύ των ατόμων 14 και 16 και η διαδικασία αυτή συνεχίζεται μέχρι να τοποθετηθούν όλα τα άτομα, τελικά σε μια κοινή ομάδα.

Agglomeration Schedule

Stage	Cluster Combined			Stage Cluster First Appears		
	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	5	18	306,948	0	0	3
2	14	16	606,953	0	0	3
3	5	14	979,574	1	2	5
4	3	13	997,283	0	0	7
5	5	15	1014,731	3	0	6
6	5	11	1026,683	5	0	8
7	3	6	1093,673	4	0	9
8	4	5	1204,050	0	6	9
9	3	4	1212,127	7	8	10

10	3	20	1343,337	9	0	11
11	3	10	1771,868	10	0	12
12	2	3	1845,683	0	11	13
13	2	12	1883,844	12	0	15
14	8	9	1939,203	0	0	15
15	2	8	1943,057	13	14	16
16	2	17	2297,903	15	0	17
17	2	7	2724,506	16	0	19
18	1	19	3416,463	0	0	19
19	1	2	3820,091	18	17	0







Στη συνέχεια φαίνεται εποπτικά η προηγούμενη διαδικασία. Στον κάθετο άξονα είναι τοποθετημένα τα άτομα και στον οριζόντιο οι αποστάσεις. Όπως παρατηρείται στην αρχή σχηματίζει ομάδα το άτομο 5 με το άτομο 18, στη συνέχεια το 14 με το 16. Έπειτα και τα τέσσερα αυτά άτομα κάνουν όλα μαζί μια ομάδα. Στη συνέχεια το 3 με το 13 και σε επόμενο στάδιο το 15 με το 11 σχηματίζουν ομάδα. Καθώς αυξάνεται η απόσταση η ομάδα (15,11) ενώνεται με αυτή των (5, 18, 14, 16) κ.ο.κ. Τελικά φαίνονται δύο ομάδες μία μεταξύ των 1 και 19 και μία μεταξύ των υπολοίπων οι οποίες στο τέλος ενώνονται σε μια και μοναδική. Το δεντροδιαγραμμα παρακάτω βοηθάει στην κατανόηση υλοποίησης του αλγόριθμου.



K-MEANS ΜΕΘΟΔΟΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Στον πρόβλημα που έχει παρουσιαστεί στην αρχή του κεφαλαίου θα το χρησιμοποιήσουμε ως βάση για την υλοποίηση της συσταδοποίησης με την μέθοδο k-means. Τα δεδομένα τώρα που θα χρησιμοποιήσουμε είναι τα παρακάτω όπως αυτά φαίνονται στο αρχικό φύλλο του spss.

Αρχικά, εισάγουμε τα δεδομένα στο spss και ονομάζουμε και δηλώνουμε τις μεταβλητές.

 PATIENT SID	 B1	 B2	 E1	 E3	
1	372,00	292,00	13826,00	15223,00	
2	260,00	151,00	7729,00	16088,00	
3	193,00	193,00	7173,00	11729,00	
4	272,00	200,00	7535,00	15043,00	
5	234,00	161,00	5830,00	12958,00	
6	213,00	139,00	6660,00	12527,00	
7	205,00	151,00	5575,00	8178,00	
8	210,00	124,00	4153,00	10498,00	
9	214,00	162,00	5752,00	10785,00	
10	238,00	187,00	6579,00	11141,00	
11	230,00	176,00	7005,00	13997,00	
12	163,00	185,00	7356,00	16875,00	
13	188,00	164,00	7826,00	11009,00	
14	194,00	173,00	6148,00	13955,00	

Στην συνέχεια επιλέγοντας

Analyze → classify → k-means clustering

Θα εμφανιστούν στο output τα αποτελέσματα.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	Df	Mean Square	Df		
B1	11631,000	2	913,766	11	12,729	,001
B2	7353,250	2	519,175	11	14,163	,001
E1	25324230,080	2	1053605,342	11	24,036	,000
E3	28028228,620	2	2082848,199	11	13,457	,001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Στον πίνακα ANOVA θα πρέπει να εξεταστεί αν είναι τα δεδομένα στατιστικά σημαντικά, (δηλαδή $p\text{-value} < 0.05$) έτσι ώστε να εξασφαλιστεί η σωστή συσταδοποίηση.

Αν το p-value σε κάποια από τις μεταβλητές ήταν μεγαλύτερο από το όριο του 0.05 τότε θα πρέπει να αλλάχτει ο προαπαιτούμενος αριθμός συστάδων (cluster) που έχει βάλει ο χρήστης. Στην περίπτωση του παραδειγματος που γίνεται η ανάλυση οι μεταβλητές είναι στατιστικά σημαντικές.

Number of Cases in each Cluster

Cluster	1	1,000
	2	7,000
	3	6,000
Valid		14,000
Missing		,000

Στον παραπάνω πίνακα φαίνεται πόσα στοιχεία ή αντικείμενα έχει η κάθε συστάδα, δηλαδή η πρώτη συστάδα έχει ένα στοιχείο, η δεύτερη έχει επτά και τέλος η τρίτη έχει έξι στοιχεία.

ID	B1	B2	E1	E3	QCL_1	QCL_2
1,00	372,00	292,00	13826,00	15223,00	1	,00000
2,00	260,00	151,00	7729,00	16088,00	2	1801,48708
3,00	193,00	193,00	7173,00	11729,00	3	1539,09231
4,00	272,00	200,00	7535,00	15043,00	2	846,75801
5,00	234,00	161,00	5830,00	12958,00	2	1867,21951
6,00	213,00	139,00	6660,00	12527,00	2	1979,08727
7,00	205,00	151,00	5575,00	8178,00	3	2453,53258
8,00	210,00	124,00	4153,00	10498,00	3	2024,57003
9,00	214,00	162,00	5752,00	10785,00	3	481,90574
10,00	238,00	187,00	6579,00	11141,00	3	710,66035
11,00	230,00	176,00	7005,00	13997,00	2	507,08098
12,00	163,00	185,00	7356,00	16875,00	2	2428,18606
13,00	188,00	164,00	7826,00	11009,00	3	1710,67408
14,00	194,00	173,00	6148,00	13955,00	2	920,16007

Έχοντας επιλέξει την επιλογή να αποθηκευτούν σε ποια συστάδα έχει κατηγοριοποιηθεί το κάθε αντικείμενο και ποια είναι η απόσταση αυτού από την το κέντρο της συστάδας, στην αρχική σελίδα του spss έχουμε τις νέες δυο στήλες (QCL_1, QCL_2) όπως φαίνεται και παραπάνω.

Επομένως, στην πρώτη συστάδα βρίσκονται οι ασθενείς: 1

Στην δεύτερη συστάδα βρίσκονται οι ασθενείς: 2, 4, 5, 6, 11, 12, 14

Στην Τρίτη συστάδα βρίσκονται οι ασθενείς: 3, 7, 8, 9, 10, 13

ΣΥΖΗΤΗΣΗ

Από τα παραπάνω έχει γίνει σαφές πως η ιεραρχική μέθοδος συσταδοποίησης είναι ευρέως διαδεδομένη και σε αυτό συνηγορεί ότι ο αλγόριθμος που την διέπει είναι σχετικά απλός και αρκετά εύκολος. Επιπλέον, δεν απαιτείται στους ιεραρχικούς αλγορίθμους αρχικός προσδιορισμός των συστάδων που θα δημιουργηθούν. Παρόλα αυτά, δεν μπορεί αυτή η μέθοδος συσταδοποίησης να ανταπεξέλθει επαρκώς στα θορυβώδη σημεία και στις μεγάλες απαιτήσεις αποθήκευσης, σε αντίθεση με την μέθοδο k-means που ανήκει στους διαμεριστικούς αλγορίθμους. Επίσης, πάροτι ο k-means αλγόριθμοι είναι αρκετά αποδοτικός χρειάζεται για την υλοποίηση του προσδιορισμός του αριθμού των συστάδων γεγονός που σε πολλές περιπτώσεις είναι δύσκολος. Ένα άλλο αρνητικό σημείο για αυτήν την κατηγορία αλγορίθμου είναι ότι η χρήση του περιορίζεται μόνο σε σφαιρικά δεδομένα. Αντίθετα, στον αλγόριθμο συσταδοποίησης DBSCAN που ανήκει στους αλγορίθμους πυκνότητας δεν απαιτείται κάποιο συγκεκριμένο σχήμα στα δεδομένα. Επίσης, για την υλοποίηση του απαιτείται ο προσδιορισμός δύο παραμέτρων (MinPts, ϵ) που όμως δεν είναι δύσκολος. Ένα μειονέκτημα της μεθόδου όμως είναι ότι η ποιότητα των αποτελεσμάτων εξαρτάται από τη μετρική απόστασης που θα χρησιμοποιηθεί. Τέλος, έγινε αναφορά και σε δύο ακόμα μεθόδους την GRID-BASED διαδικασία και την MODEL-BASED. Το κύριο πλεονέκτημα των GRID-BASED μεθόδων είναι πως η ταχύτητά τους εξαρτάται μόνο στην ανάλυση του πλέγματος και όχι στο μέγεθος της βάσης δεδομένων. Τέτοιες μέθοδοι είναι κατάλληλες περισσότερο για σύνολα δεδομένων υψηλής πυκνότητας με μεγάλο πλήθος αντικειμένων σε μικρό χώρο. Οι MODEL-BASED αλγόριθμοι προσπαθούν να βελτιστοποιήσουν το ταίριασμα μεταξύ των δεδομένων και ενός μαθηματικού μοντέλου που θα τα περιγράφει.

Από τα παραπάνω γίνεται σαφές ότι κάθε μέθοδος έχει τα πλεονεκτήματα και τα μειονεκτήματα της. Θα πρέπει να τονισθεί ότι δεν υπάρχει σωστή και λάθος μέθοδος συσταδοποίησης αλλά αυτό εξαρτάται από το είδος και το μέγεθος των δεδομένων που πρέπει να αναλυθούν και τι συμπεράσματα αναζητά ο ερευνητής. Είναι επίσης πολύ σημαντικό να σημειωθεί πως η συσταδοποίηση δεν αντιπροσωπεύει πάντα την πραγματικότητα. Σκοπός είναι να βρεθεί η καλύτερη μέθοδος συσταδοποίησης έτσι ώστε να γίνει η καλύτερη ανάλυση και τα δεδομένα να μπορούν να αξιολογήσουν το γενικό σύνολο.

Η ΚΑΤΑΛΛΗΛΗ ΜΕΘΟΔΟΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ΘΑ ΠΡΕΠΕΙ ΝΑ

1. Χειρίζεται ικανοποιητικά πολλά δεδομένα
2. Αναγνωρίζει ακραία σημεία καθώς και θόρυβο
3. Αναγνωρίζει συστάδες ποικίλου μεγέθους, πυκνότητας και σχήματος.
4. Χειρίζεται δεδομένα πολλών διαστάσεων

5. Παρέχει συστηματικό τρόπο για την αναγνώριση του αριθμού των συστάδων
6. Απαιτεί μικρό βαθμό παραμετροποίησης και μικρό αριθμό παραμέτρων

Σύμφωνα με τις μεθόδους που έχουν αναλυθεί στα προηγούμενα κεφάλαια, παρακάτω παρουσιάζεται ένας πίνακας που αναλύει ποια από τα έξι χαρακτηριστικά έχει η κάθε μέθοδος. Καμία μέθοδος όπως είναι προφανές δεν έχει και τα έξι χαρακτηριστικά. Αν αυτό συνέβαινε τότε θα είχε βρεθεί η καταλληλότερη μέθοδος συσταδοποίησης που θα ικανοποιούσε όλα τα δεδομένα.

Μέθοδοι	1	2	3	4	5	6
Partitioning	Yes/No	No	No	Yes	No	No
Hierarchical	No	No	Yes/No	Yes	Yes/No	Yes/No
Density-based	Yes	Yes	Yes	No	Yes/No	Yes/No
Grid based	Yes	Yes	Yes/No	No	Yes/No	Yes/No
Model-based	No	Yes	Yes/No	No	Yes	Yes/No

Κλιμάκωση σε μεγάλες βάσεις δεδομένων

Οι Partitioning, Hierarchical και Model-based μέθοδοι δεν κλιμακώνονται ικανοποιητικά όταν ο αριθμός των δεδομένων είναι πολύ μεγάλος. Οι Grid based μέθοδοι είναι πολύ αποτελεσματικές στις κλιμάκωση επειδή συμπυκνώνουν τα δεδομένα, με ενδεχόμενο φυσικά να χαθούν πληροφορίες. Οι Density-based μέθοδοι κλιμακώνονται ικανοποιητικά και ανταγωνίζονται τις Grid based δίχως το ρίσκο της συμπύκνωσης των δεδομένων. Παρότι δεν φαίνεται στον πίνακα σύγκρισης ο DBSCAN είναι δυνατόν να χειριστεί αποτελεσματικά δεδομένα με περισσότερες από 20 διαστάσεις και η επιλογή της παραμέτρου Eps μπορεί να προεκτιμηθεί από τις μέσες αποστάσεις των σημείων.

ΚΕΦΑΛΑΙΟ 7

ΑΝΑΦΟΡΕΣ

1. Guonjun Gan, Chaoqun Ma, Jianhong Wu editors, ‘Data clustering Theory Algorithms and Applications’ Siam Publications, 2007
2. Jiawei Han and Micheline Kamber “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, August 2000.
3. M.H. Dunham, “Data Mining introductory and advanced topics” Prentice Hall 2004.
4. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, editors. ‘Introduction to Data Mining’, Pearson International Edition, pg 487- 2016.
5. M. S. Aldenderfer and R. K. Blashfield. Cluster analysis. Sage publication, Los Angeles, 1985
6. R. Ng and J. Han “Efficient and effective clustering method for spatial data mining”, in Proc. Conf. on Very Large Data Bases , pg 144-155, 1994.
7. Tian Zhang, Raghu Ramakrishnan, Miron Livny, editors. ‘BIRCH: A New Data Clustering Algorithm and Its Applications, Springer US 1997.
8. Hierarchical Cluster Analysis. Ανακτήθηκε στις 27 Νοεμβρίου 2015, από: <http://www.r-tutor.com/gpu-computing/clustering/hierarchical-cluster-analysis>
9. Ζιντζαρας, 2016 ασκήσεις στο μάθημα στα προηγμένα στατιστικά πακέτα.